**Practical Application Assignment 17.1**

**What Makes a Successful Direct Marketing Campaign**

The practical application assignment targets to compare the performance of the classifiers of Logistic Regression, K Nearest Neighbor, Decision Trees, and Support Vector Machines.

1. Data Understanding

The dataset is the direct marketing campaigns (phone calls) of a Portuguese banking institution. The goal of the classification is to predict if the client will subscribe a term deposit (variable y).

The data were collected from 2008 to 2013, in total of 52944 phone contacts. It includes the effects of the financial crisis period. The dataset contains a large set of 150 features which refers to bank client, product and social-economic attributes. The result is a binary unsuccessful or successful contact of the telemarketing phone calls to sell the long-term deposits.

The dataset is unbalanced, as only 6557 (12.38%) records are related with successes as shown in the pie visualization plot below.

1. Business Objective

The Business Objective is to apply data mining technique by applying machine learning models to predict the success of telemarketing calls for selling bank long-term deposits.

By analyzing the features related with bank client, product and social-economic attributes, we can optimize for targeting telemarketing calls in order to increase bank profits and reduce costs.

1. Features Engineering

The feature engineering phase starts with getting familiar with the data, discovering first insights into the data, and to detecting important subsets to form hypotheses.

For example, before dropping the 'age' feature, I will check its influence with reference to marital on the success of the calls. I generated the histogram of age stratified by marital and y as shown below which shows the impact is age is relatively uniform and may be dropped.

A graph of a bar graph

Description automatically generated with medium confidence

After exploring the dataset, I found that many features are categorical values in object type and should be encoded to numerical values. These features include 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'day\_of\_week', 'poutcome', as well as the final outcome of 'y'.

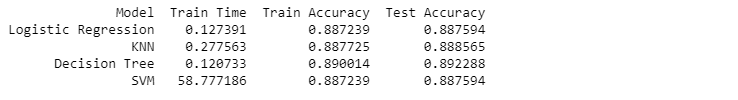
To form the data for modeling, only the bank information features (columns 1 - 7) for model fitting are kept. I used LabelEncoder() to the encode the column to transform the features from categorical values to numerical values.

1. Modeling

For modelling, I first created the baseline model using DummyClassifier. This will act as a comparison to the actual models.

After that, I created the Logistic Regression, KNN, Decision Tree, and SVM models one-by-one using the default settings for each of the models. I then conducted the comparisons of these models by fitting, scoring and calculating the fit time of each of the models.

Based on the outcome, I generated the performance matric as shown below.



For visualization, I created bar plot of the model performance matric comparison as shown below.

A graph with blue bars

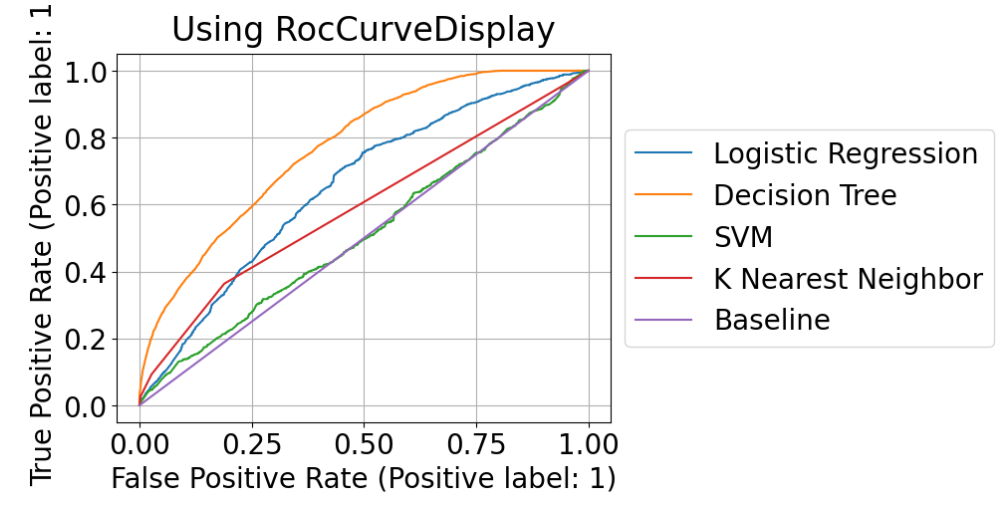
Description automatically generated with medium confidence

1. Model Improvement

For model improvement, I first examined the feature engineering again as it is important to increase the accuracy of model classification.

Whether the gender feature should be kept? Keeping the gender feature would be beneficial for model improvement. This is because during direct marketing calls, gender influence (male/female) which are related to three categories, including the gender of the banking agent, client and client-agent difference, can impact the outcome of successful and unsuccessful calls. Therefore, gender feature can be kept in order to increase the accuracy of the classification.

For hyperparameter tuning and grid search, I used RocCurveDisplay to analyze the performance of these five classification models at all classification thresholds. The visualization plot is shown below.



From the result of RocCurveDisplay, it is determined that the Decision Tree model can provide the best performance of classification.

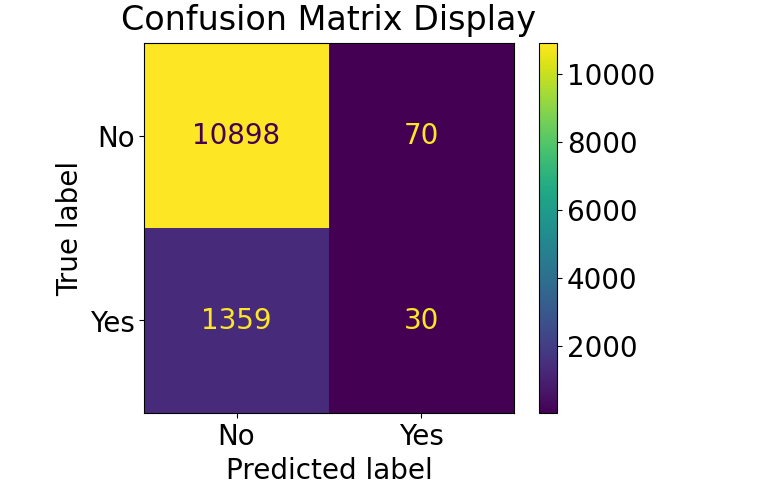
After that, I performed the search for the best max\_depther of the Decision Tree model. The training and test accuracy are calculated and the visualization plot is created as shown below.

A graph with red and green lines

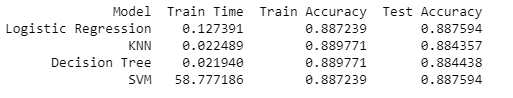
Description automatically generated

From the Decision Tree training and test data accuracy curve above, the best max\_depth is chosen accordingly.

Next, I calculated the confusion matrix of the best model and created the visualization plot as shown below.



Finally, I recalculated the performance matric for the Logistic Regression, KNN, Decision Tree of max\_depth = 11 and SVM as shown below.



1. Conclusion

The findings of this study can be summarized as the followings:

(1) The result from Confusion Matrix shows that using the best model created, the true unsuccessful rate is quite high, meaning it is easier for the caller of telemarketing campaigns to tell which client may decline to subscribe a term deposit, while the true successful rate is relatively low, meaning it is harder to determine which client may accept to subscribe a term deposit.

(2) From RocCurveDisplay, the Decision Tree's ROC curve is closer to the upper left corner of the graph, which means it provides higher the accuracy of the test data. It was further identified that max\_depth = 11 can produce better accuracy. However, it may run into overfitting if higher value of max\_depth is chosen.

(3) The Logistic Regression, KNN, Decision Tree and SVM models produce relatively similar train accuracy and test accuracy for this dataset. Meanwhile, the SVM takes significantly more train time to complete as compared to that of the other three models.

1. Next Steps

For Next steps and recommendations, I prefer to explore further in these two areas:

(1) Understand if including more features in the model fitting would increase the accuracy of classification. Instead of using only the bank information features (columns 1 - 7), we could include other features such as 'age', product and social-economic attributes to find out their impact on the classification accuracy.

(2) Understand why the SVM is so slow as compared to Logistic Regression, KNN and Decision Tree models. When performing the training, the SVM is not conducted incrementally. Instead, it requires to use the entire dataset to be trained all at once. As a result, if there are more data points, it's going to run long time to complete.

In order to speed up the non-linear kernel of SVM, it is useful to adopt cross-validation or grid search to find the optimal values for a well-tuned model that will converge faster, or use SGDClassifier and kernel approximator like Nystroem for better hyperparameter tuning. These steps can be considered for further investigation.